データ自動再配置ストレージによるレスポンスタイム改善と柔軟な運用の実現

丸山 一貴[1,2]　山原 陽一[3]　関谷 貴之[2]

概要：東京大学情報基盤センターは，2012 年 3 月に教育用計算機システムを更新し，ECCS2012 を導入した．1,300 台以上の iMac を使う 40,000 ユーザに対するホームディレクトリサービス（以下，ホームサービスという）と，60,000 ユーザに対するメールサービスのためにストレージが必要である．従来システムでは，ホームサービスとメールサービスは独立しており，一方の空き容量を他方に融通することができなかった．また，ホームサービスでは授業利用時のアクセスピークにおいてパフォーマンス不足によりレスポンスの大幅な低下を招いていた．
これらを解決するため，SSD/FC/SATA の各ドライブを混在して 3 つの階層として扱い，アクセス頻度の高いデータほど高速なドライブに配置するハイエンドストレージを導入した．2013 年 4 月から 6 月までの稼働データを収集し，各階層の使用状況を追跡した．加えて，従来型の「単一種類のドライブによるディスクアレイ」を選択した場合に比べて，レスポンスタイムが約 0.76 倍に改善したことをシミュレーションにより明らかにした．

キーワード：統合型ストレージ，自動再配置，レスポンスタイム，教育用計算機システム

Benefits of integrated storage with automatic tiering for home and mail services

Abstract: Information Technology Center, The University of Tokyo currently provides Educational Campuswide Computing System 2012 (ECCS2012), which was launched in March 2012, to 60,000 users at our university. ECCS2012 includes two network file services: (1) a file service to 1,300+ iMacs for 40,000 users (home directory service), and (2) mail servers for 60,000 users (mail service). In our prior system, ECCS2008, these services could not share free spaces of their drives because of their completely separated storages. The home directory service had a performance problem at a peak time of disk access
In ECCS2012, EMC Symmetrix VMAX, a single, high-end storage with automatic tiering of SSD/FC/SATA, resolved the problems above. It locates frequently accessed data to faster drives and enables us to get better balance between costs and I/O performance. We have measured the response time and collected the statistics for three months. We have also simulated a single tier, the same size array of FC drives in order to show that the multiple tiers of drives have better performance in the response time and show that the average response time are reduced to 0.76 times approximately.

Keywords: Unified storage, automatic tiering, response time, educational computer system.

1. 背景

東京大学情報基盤センターの情報メディア教育部門では，東京大学構成員の情報基盤として教育用計算機システム (Educational Campuswide Computing System，以下 ECCS という) を開発・管理・運用している．同システムは講義等で利用する PC 端末 [1], [2]，PC 端末及びメールホスティングサービスのユーザが使用するメールシステム，これらに付随する認証及びユーザ管理システム [3]，ファイルサーバ，プリンタ [4], [5]，各種サーバを動作させる仮想

[1] 明星大学 情報学部
School of Information Science, Meisei University
[2] 東京大学 情報基盤センター
Information Technology Center, The University of Tokyo
[3] EMC ジャパン株式会社
EMC Japan K.K.

マシン環境基盤，ネットワーク装置からなる．当センターでは 2012 年 3 月に同システムを更新し，新システム（以下，ECCS2012 という）の運用を開始している．

ECCS2012 のファイルサーバは，主に 3 つのサービスを提供している：(1) 1,300 台以上の iMac を使う 40,000 ユーザに対してファイル保存領域を提供するホームディレクトリサービス（以下，ホームサービスという），(2) 合計 60,000 ユーザが利用するメールサーバと連携してメール保存領域を提供するサービス（以下，メールサービスという），(3) 仮想マシン環境基盤に仮想ディスク領域を提供するサービス（以下，SAN サービスという）である．これらのうち (1) と (2) はストレージ容量の大部分を占めており，また授業及びメールシステムで利用されることから，大きな容量を確保しつつレスポンスタイムを維持することが必要である．加えて，各利用者の利用状況やサービスの申し込み状況によっては，一方のサービスの空き容量を他方に融通することが必要になることも想定される．

我々はこれらの課題に対して，2 つの方法により対処した．第 1 は，SSD，FC HDD（以下，FC という），SATA HDD（以下，SATA という）という 3 種類のドライブを混在させて 3 つの階層を構成し，ブロックの利用頻度に応じて適切な階層に再配置する技術を利用したことである．第 2 は，ホームサービスとメールサービスで利用する領域を単一のストレージプールとして構成することで空き容量を共通化し，容量が不足したサービスのボリュームに動的に追加する構成としたことである．本論文ではこれらの具体的な構成法について述べるとともに，従来型と同様に単一種類のドライブのみを利用した場合と比べ，レスポンスタイムがどの程度改善されたかをシミュレーションにより示す．以下，第 2 章では前システムにおけるストレージの構成と問題点を，第 3 章で ECCS2012 におけるストレージの特徴と構成を，第 4 章では 2013 年 3 月から 6 月にかけての利用状況に基づいてストレージへの負荷の特徴と傾向を述べる．第 5 章では，階層化によりレスポンスタイムが改善していることを，シミュレーションにより示す．第 6 章に関連研究を，第 7 章でまとめと今後の課題を述べる．

2. ECCS2008 におけるストレージの問題点

2008 年 3 月に稼働を開始した前システム（以下，ECCS2008 という）におけるストレージの構成を，**図 1** に示す．図中でグレーにしている箇所は，ECCS2012 におけるホームサービス及びメールサービスに移行したストレージを表している．

メールサービスのために，アプライアンス方式と汎用サーバ方式で合計 5 つのストレージが存在していた．メールホスティングサービスは，学内で独自にメールサーバを運用していた部局 *1 が，サーバ運用が困難になって利用を申し込むケースが多く，希望するディスク容量も様々であり，利用予測が困難である．2 方式分のユーザサポートコストだけでなく，ストレージが分割されていたことから，2 つの方式間で空き容量を融通することができないという懸念も問題となっていた．以上のことから，ECCS2012 においては，メールサービスを 1 方式に統一することと，ストレージを統合することは重要な課題となっていた．

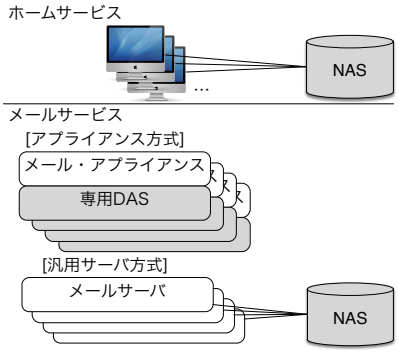

図 1 ECCS2008 のストレージ構成
Fig. 1 Storages of ECCS2008.

ホームサービスでは性能面の問題があり，授業利用時のアクセスピークにおいてパフォーマンス不足によりレスポンスの大幅な低下を招いていた．一方で，ストレージ利用量は PC 端末のユーザ数が支配的な要因である．ユーザの大部分は学生であり，学生定員はほとんど変化しないため，利用量が予測しやすいサービスであると言える．従って，ホームサービスとメールサービスで空き容量を共通化し，いずれかのサービスで容量が不足したとき機動的に空き容量を追加することができれば，余剰の容量と運用上の懸念を共に軽減できると考えた．性能と容量という 2 つの問題を解決すべく，ECCS2012 を設計した．

3. ECCS2012 におけるストレージの構成

ECCS2012 ではホームサービスとメールサービスを 1 つのストレージに統合することを前提とした．ホームサービスは授業時間中の一斉ログイン等で負荷がかかる場合にも安定したレスポンスタイムが実現できる性能を，メールサービスは 24 時間 365 日にわたってできる限り停止しない高信頼性を必要としていた．従って，ハイエンドストレージと NAS ヘッドの組み合わせでこれらサービスを実現すると共に，SAN サービスも統合する設計とした．ECCS2012 はストレージとして EMC 社製 Symmetrix VMAX を，NAS ヘッド *2 として同 VNX VG8 を使用している．

3.1 Symmetrix VMAX

Symmetrix VMAX は 3 種類のドライブを混在して搭載

*1 学科や研究室という単位で，独自のドメイン名を利用して運用している．

*2 FC 等の SAN プロトコルと NFS 等の NAS プロトコルとの変換を行う装置．

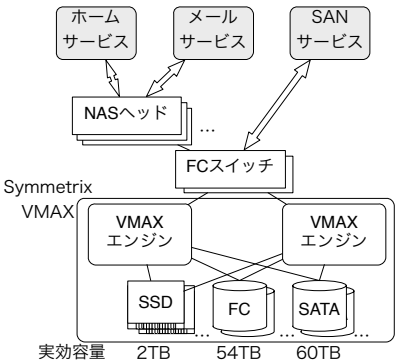

図 2 Symmetrix VMAX の内部構成
Fig. 2 Details of Symmetrix VMAX.

できる，FC 接続のストレージ製品である．内部にはドライブを束ねるストレージコントローラとしての VMAX エンジンを搭載している．外部から届く特定のブロックに対する読み書きのリクエストは，VMAX エンジンによって具体的なドライブ上のセクタへと変換され，物理的なアクセスが行われる．

ECCS2012 におけるストレージ及び NAS ヘッドと各サービスとの接続を**図 2** に示す．ホームサービスでは，iMac 上で OS X が動作する場合は NFS で，Windows が動作する場合は CIFS で NAS ヘッドにアクセスし，ストレージ内のデータを読み書きする．メールサービスでは，IA サーバ上の Linux が NFS で NAS ヘッドにアクセスする．SAN サービスでは NAS ヘッドを介さず，FC スイッチに直接接続する．2 台の VMAX エンジンにはそれぞれ 64GB のキャッシュ搭載しており，一方をホーム及びメールサービス用，他方を SAN サービス用として動作させている．これらは Active-Active の HA 構成であり，一方の VMAX エンジンに障害が発生した場合は他方が 2 台分の処理を行う設定である．

VMAX には 200GB の SSD，15,000rpm で 600GB の FC ディスク，7,200rpm で 2TB の SATA ディスクの 3 種類を混在して搭載しており，これらを 2 つのグループに分けて使用する．1 つは SAN サービス用のグループであり，FC ドライブを RAID 構成後の実効容量で 6TB 割り当てた．SAN サービス用のドライブは仮想マシンの仮想ディスクが配置されるドライブであり，ECCS2012 で必要な領域は十分に計画済みで使用量の上限が確定している．もう 1 つはホーム及びメールサービス用のグループであり，残りのドライブ全てを割り当て，単一のストレージプールとして構成した（第 3.2 章）．どのデータブロックをどのドライブに配置するかは当該ブロックの利用頻度に応じて VMAX エンジンが決定する（第 3.3 章）．

3.2 ストレージプールによる柔軟な運用

ホーム及びメールサービス用のドライブは，実効容量で SSD が 2TB，FC が 48TB，SATA が 60TB の合計 110TB 分あり，これらを単一のストレージプールに構成した．このプールから複数のボリュームを切り出して NAS ヘッドからマウントし，ファイルシステムを構成して NFS 及び CIFS クライアントへエクスポートする．運用開始時点では未使用領域を十分残した状態にしておき，ホーム及びメールの使用率を見定めながら，不足したサービスのボリュームに容量を追加する（ファイルシステムを拡張する）設計とした．本稿執筆時点で，ホームには合計 32TB を，メールには合計 30TB を割り当てている [*3]．

3.3 自動階層化

ホーム及びメールサービス用のドライブ群は，単一のプールであるだけでなく EMC FAST VP[6] を用いた自動階層化を有効化している．この機能は，VMAX エンジンがブロック単位でアクセス統計を取り，利用頻度の高いブロックをより高速なドライブに割り当てるものである．従って，頻繁にアクセスされるデータは SSD に割り当てられ，ほとんどアクセスされないデータは SATA に移動することとなる．この機能を利用した目的は以下の通り．

- ECCS2008 より総容量を拡張するため，低価格・大容量なドライブを活用すること．
- 大容量ドライブの導入でスピンドル数を減らし，消費電力を抑えること．
- SSD や FC といった高速なドライブを活用し，大容量ドライブによるアクセス性能低下を補うこと．

本機能により各ドライブに割り当てられる容量がどのように変化するかは第 4 章で，FC ドライブのみで構成した場合との性能上の比較については第 5 章で述べる．

4. 負荷の傾向

本章では，Symmetrix VMAX への負荷の内訳や，各階層の利用量が時間の経過に応じてどのように変化したかについて述べる．VMAX の容量及び I/O 負荷は大部分がホーム及びメールサービスによるものであり，また FAST VP による階層化も当該サービスの使用領域に対して設定した．従って，本論文で議論する対象はホーム及びメールサービスのみとする．また，ファイルサーバの負荷という観点では，NAS ヘッドと VMAX エンジンという 2 つの観測点があり得る．本論文では 3 階層の使用率や，これらが FAST VP によりどのように活用されるかに注目するため，VMAX エンジンを観測点とした．以降で示すデータは，図 2 における VMAX エンジンのうち，ホーム及びメールサービス用として設定された方で収集したものである．長期間の観測により，NAS ヘッドから来るブロックストレージへのアクセスは，その I/O サイズがほぼ固定的

[*3] これらとは別に，端末に供給するライブラリ等の領域や，端末管理システムのための領域などに約 6TB を割り当てている．

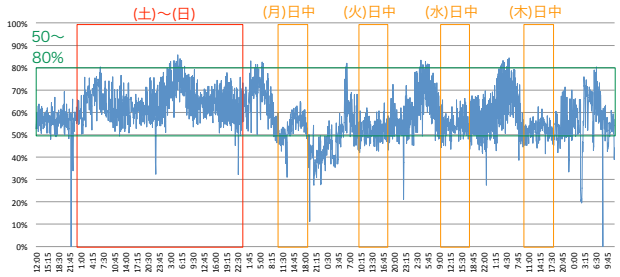

図 3 メール負荷における Write 比率
Fig. 3 Write ratio from mail servers.

に 10KByte であることが確認できた．ストレージへのアクセスは，時々刻々と変化するものであり，メールとホームディレクトリの特徴的な傾向を次に示す．

図 3 は，2013 年 3 月 22 日 (金) の夜から 2013 年 3 月 29 日 (金) の朝にかけての，負荷の大部分がメールである場合の Write 比率を示したものである．VMAX エンジンではアクセスがブロック単位となっており，ホームサービスとメールサービスいずれのボリュームに対するアクセスかは区別できないが，この時期は授業がなく，PC 端末の利用はごくわずかであり，負荷のほとんどがメールサービスに起因するものと言える．図中で緑色の矩形で示した箇所は，Write の割合が 50%から 80%の範囲を表している．平日の日中は 50%強が多いが，平日の夜間及び土日は 80%程度まで増えており，メールサービスに固有の特性を示している．メールは日中や夜間に限らず受信し続けており，Write の負荷は常にかかっているが，日中は届いたメールへの Read アクセスがあり，相対的に Write 比率が下がることになる．

ホームサービスに起因する負荷の傾向として，2013 年 4 月 1 日から 2013 年 6 月 21 日の期間における，SSD/FC/SATA の各ドライブに対して割り当てられたブロックの合計容量の推移を**図 4** に示す．上段から順に，SSD/FC/SATA の容量を示しており，上部の横線が実装容量を表している．SSD は実装容量のほぼ 100%を常に使い切っており，FAST VP により高速なドライブが活用されていることが分かる．FC は徐々に割り当て容量を増やし，SATA は容量を減らしている．集計期間は夏学期の授業期間であり，同じユーザがほぼ毎週 PC 端末を利用している．従って授業を履修しているユーザの設定ファイル群 [*4] や，授業で作成したデータファイルなどは，学期中に継続してアクセスされる傾向にあると考えられる．このため，新たなブロックは FC 上に割り当てられるものの，より低速な SATA へは移動しにくい．期間中に FC で増加した容量は約 12TB なのに対して，SATA で減少した容量は約 10TB である．新規に生成されたファイルに加えて，アクセスされるブロックが徐々により高速な FC 側へと移動していると考えられる．

[*4] OS X ではユーザ個人の設定ファイル群もホームディレクトリ配下にあり，ホームサービスにより PC 端末へ供給される．

図 4 各階層の割り当て容量の推移
Fig. 4 Trend of allocated size of each tier.

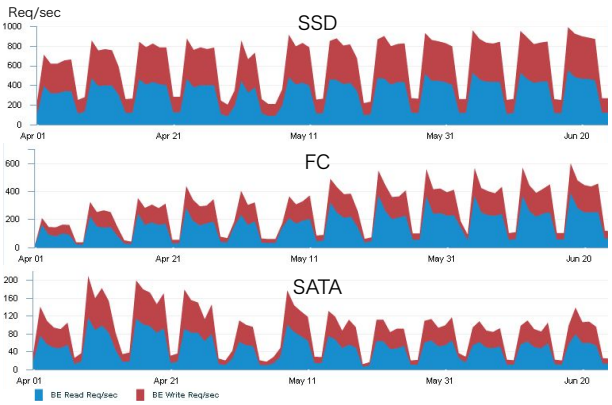

図 5 各階層の IOPS の推移
Fig. 5 Trend of IOPS of each tier.

このデータにはメールサービスも含まれるが，メールサービスでは固定的なスプール領域に対して読み書きが行われるため，階層間の移動傾向にはあまり寄与しないと考えられる．また，長期休暇において同様の傾向を調査すると，授業がなく PC 端末の利用者が少ないことから FC のデータが SATA へ移動することにより，FC 及び SATA の割り当て容量推移が逆転することが観測されている．

次に，2013 年 4 月 1 日から 2013 年 6 月 21 日の期間における，SSD/FC/SATA の各階層に対する 1 秒当たりの I/O 数（以下，IOPS という）の，1 時間ごとの平均値の推移を**図 5** に示す．特定の階層に対して発生する IOPS は，当該階層に割り当てられた容量が主たる要因となるため，集計期間における推移は図 4 と同様の傾向を示す．6 月 20 日前後で階層間を比較すると，SSD は約 850IOPS を発揮しており，約 430IOPS である FC の 2 倍近くの I/O を処理している．SATA の約 100IOPS も合わせると，全体の約 62%は SSD が処理しており，性能の観点でも SSD が有効活用されていると言える．

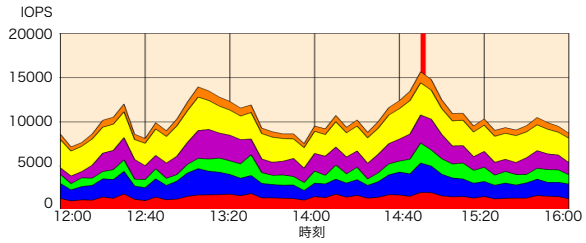

図 6 6/21 のワークロード推移
Fig. 6 Trend of workloads on 6/21.

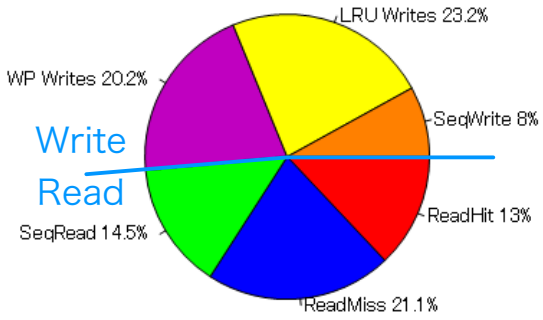

図 7 6/21 14:50 のワークロード詳細
Fig. 7 Details of the workload at 14:50 on 6/21.

5. シミュレーション

第 4 章では，FAST VP の機能により，高速だが高価で小容量の SSD から，低速だが安価で大容量の SATA までを混在して活用している傾向を示した．本章では FAST VP を利用せず，エンタープライズのファイルサーバで標準的な FC 接続のディスクのみで構成した場合に，性能差がどの程度になるかをシミュレーションによって試算する．ユーザの体感には待ち時間が最も大きな影響を及ぼすと考え，比較する性能指標はレスポンスタイムとする．

5.1 比較に用いるワークロード

図 6 は 2013 年 6 月 21 日 12:00 から 16:00 に発生した，ホーム及びメールサービスの IOPS 内訳（以下，ワークロードという）の，5 分ごとの平均値の推移を示したものである．グラフの各色については後述する．図中の赤い縦線で示した 14:50 時点でピークが発生しているが，この時間は 4 限の授業が始まったばかりであり，100 人前後が利用する大演習室が複数，連続して利用され，多人数のログアウトとログインが交錯してホームサービスの負荷が高くなる．このとき IOPS は 15,675，スループットは 177MB/s，平均 I/O サイズは 11.56KB であった [*5]．シミュレーションには，このピーク時におけるワークロードを用いる．

図 7 は 14:50 時点のワークロード詳細を示しており，図中の色は図 6 と対応している．青い線で区切った下側が Read，上側が Write を表しており，Read は合計で 48.6%，Write は合計で 51.4%である．Read/Write の詳細を**表 1** に示す．これらのうち，キャッシュに対する読み書きはストレージとして最も高速な応答である．従って，レスポンスタイムに悪影響を与えるのは，データをドライブまで読みに行った上で応答を返す ReadMiss のみであると言える．ReadMiss 以外でドライブにアクセスするものに SeqRead と LRU Writes がある．いずれもドライブの負荷を上げることになるので，結果として全体的なレスポンスタイムを増大させる可能性はあるが，そのオペレーションの完了を待っているユーザプロセスがあるわけではなく，レスポンスタイムの算出においては ReadMiss に比べて無視できるものと判断した．

表 1 Read/Write の内訳
Table 1 Details of read/write.

種別	意味
ReadHit	不連続アドレスへの Random Read で，VMAX エンジンのキャッシュにヒットしたアクセス
ReadMiss	不連続アドレスへの Random Read で，キャッシュにヒットせずドライブからデータを読み出したアクセス
SeqRead	連続アドレスへの Sequential Read で，プリフェッチにより予めドライブからデータを読み出してキャッシュから返すアクセス
WP Writes	不連続アドレスへの Random Write で，キャッシュに書いた時点で Ack を返すアクセス
SeqWrite	連続アドレスへの Sequential Write で，キャッシュに書き込んだ時点で Ack を返すアクセス
LRU Writes	キャッシュからドライブへの Write で，対象データは LRU により決定される

5.2 比較対象のレスポンスタイムの推定

比較対象として，ECCS2012 で実装したのと同じ容量を**表 2** のドライブを用いて構成した場合を想定する．なお，ECCS2012 では SSD のみ RAID5 とし，FC 及び SATA は表 2 の構成と同じく 6+2 の RAID6 を構成している．14:50 時点の IOPS が図 7 の内訳でこれら各構成のドライブ群にかかったときの平均的なレスポンスタイムを推定する [*6]．

ドライブは使用率に応じてレスポンスタイムが変化するため，まず使用率を推定する．シミュレーションツール SymmMerge を用いると，構成 1 ではドライブの使用率が 55%，構成 2 では 72%と推定される．SymmMerge は，導入システムで想定される負荷に対して，検討している

[*5] ECCS2008 で測定した際のピークは，10 分平均で 11,237IOPS であった．

[*6] ECCS2008 のストレージはレスポンスタイムの計測機能がなく，測定できていない．このため，ECCS2008 相当と考えられる構成と比較することで，FAST VP の有用性を検証している．

表 2 FC のみによる比較対象のドライブ構成
Table 2 2 configurations with only FC drives.

名称	詳細
構成 1	FC 接続 15,000rpm 600GB のドライブを 280 台，6+2 の RAID6 で構成
構成 2	FC 接続 10,000rpm 600GB のドライブを 280 台，6+2 の RAID6 で構成

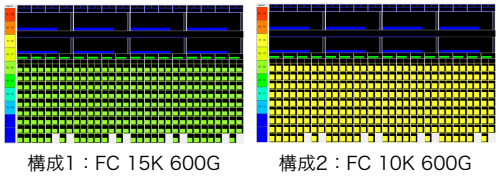

図 8 SymmMerge による算出結果の可視化
Fig. 8 Visualization of outputs by SymmMerge for FC-only configurations.

Symmetrix の構成がどのような性能を発揮するか，シミュレーションしてドライブごとの使用率を示すツールである [7]．複数の典型的な構成例を実機でベンチマーク試験した結果に基づいて，入力された構成の場合に想定される使用率を算出する．現時点では EMC 社内で有資格者のみが使用できるツールであり詳細な公開情報はないが，表 2 の構成に対するシミュレーション結果は前述の通りとなり，**図 8** のように可視化される．

次に，この使用率に基づいた平均レスポンスタイムを推定する．構成 1 及び 2 のドライブの，レスポンスタイムのベース値（使用率 0%でのレスポンスタイム）を T_S，ドライブの使用率を R_B とすると，リトルの法則から平均レスポンスタイム T_R は $T_R = T_S/(1 - R_B)$ により得られる [8]．ここで，T_S は Read の平均シーク時間と平均回転待ち時間の合計である．平均シーク時間は構成 1 が 3.4ms，構成 2 が 3.8ms である [9]．平均回転待ち時間はドライブの回転数で決まり，15,000rpm では 2ms，10,000rpm では 3ms であるから，構成 1 及び 2 の T_S はそれぞれ 5.4ms と 6.8ms である．以上より，構成 1 では ReadMiss 時のレスポンスタイムが 12ms，構成 2 では 24.3ms と算出される．この値と ECCS2012 の構成での値を比較することで，FAST VP の効果を測定する．

5.3 FAST VP におけるレスポンスタイムの推定

次に，ECCS2012 での ReadMiss によるレスポンスタイムを求める．ECCS2012 の構成では，各ドライブごとのレスポンスタイムは計測されているが，ReadMiss に限定したレスポンスタイムは測定できない．従って，以下の手順により ReadMiss 時のレスポンスタイムを推定する．

(1) 図 7 の内訳から ReadMiss に該当する Read 数を算出
(2) 階層ごとの Read レスポンスタイムを確認

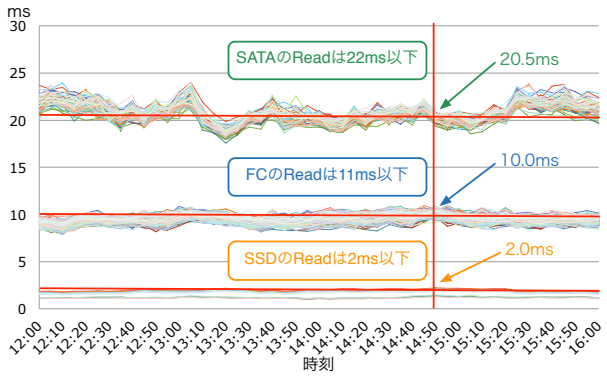

図 9 ドライブ別の Read レスポンスタイム
Fig. 9 Detailed response time of Read operations of each drive.

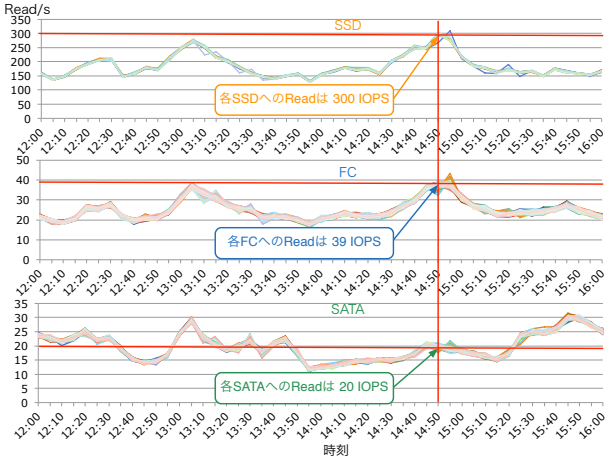

図 10 ドライブ別の Read 数
Fig. 10 Details of Read operations of each drive.

(3) 階層ごとの，ドライブあたりの Read 数を確認
(4) Read 負荷が階層ごとに分配された割合を算出し，ReadMiss に伴うレスポンスタイムを算出

手順 1 として，図 7 の内訳に基づいて 15,675IOPS を分配すると，Read 全体で 7,618IOPS，うち ReadMiss は 3,307IOPS であり，Read 全体における ReadMiss の割合は約 43.4%となる．

手順 2 として，階層ごとの Read レスポンスタイムを求める．VMAX ではドライブごとのレスポンスタイムが記録されており，14:50 時点での平均的な値は SSD が約 2ms，FC は約 10ms，SATA は約 20.5ms である（**図 9**）．

手順 3 として，階層ごとの，ドライブあたりの Read 数を求める．これも VMAX に記録されているデータを用いると，平均的な値は SSD が 300IOPS，FC が 39IOPS，SATA が 20IOPS と読み取れる（**図 10**）．

手順 4 として，まず階層ごとの Read 数を求める．手順 3 の結果に各階層のドライブ数を掛け合わせればよく，SSD は 16 本，FC は 120 本，SATA は 48 本より，各階層の合計 Read 数はそれぞれ 4,800IOPS，4,680IOPS，960IOPS となる．次に，手順 1 で算出した ReadMiss の 3,307IOPS

がこの割合で各階層にかかったと仮定すると，各階層の ReadMiss はそれぞれ 1,520IOPS，1,482IOPS，304IOPS となる．続いて，手順 2 で求めた各階層の平均レスポンスタイムを掛け合わせると，各階層の合計レスポンスタイムは 3,040ms，14,820ms，6,232ms となる．これらの合計である 24,092ms を，ReadMiss の 3,307IOPS で割って得られる約 7.3ms が，1 回の ReadMiss 当たりの平均的なレスポンスタイムと推定される．

ReadMiss のレスポンスタイムが得られたことから，図 7 の内訳と組み合わせることで，ワークロード全体に対する平均レスポンスタイムを算出したものが**表 3** である．まず，図 7 の内訳に従って，ReadMiss，その他の Read，Write の 3 つに IOPS を分配する．次に，各構成におけるレスポンスタイム欄を埋める．ReadMiss におけるレスポンスタイムの推定値は既に述べた通りである．ReadHit，SeqRead 及び Write はいずれも VMAX エンジン搭載のキャッシュに対するアクセスであるが，ここでは近似値として，SSD 階層における平均レスポンスタイムである 2ms を用いる（第 5.4 章で議論する）．種別ごとの IOPS にレスポンスタイムを掛け合わせることで合計時間を求め，構成ごとの合計時間の和を求める．例えば，FAST VP では 48,877ms となっている．これが 15,675IOPS に伴うレスポンスタイムの合計と考えられるので，IOPS で割ることにより，I/O 当たりの平均レスポンスタイムとして 3.12ms を得る．構成 1 及び 2 についても同様に求めると，FAST VP のレスポンスタイムを 1 としたとき，構成 1 は 1.32 倍，構成 2 は 2.15 倍となった．以上より，容量を同一とした場合，単一のドライブのみで構成するよりも，高速・小容量の SSD と低速・大容量の SATA ディスクを FAST VP によって統合する方が有利な性能を得られると言える．

5.4 妥当性への脅威

算出した平均レスポンスタイムの妥当性に対する脅威として，2 つの点を議論する．

第 1 は，図 9 及び図 10 の値に FAST VP による階層間データ移動の負荷が含まれる点である．VMAX エンジン内では I/O がキューにより処理されており，NAS ヘッド側からの Read アクセス（以下，上位側 Read という）は，データ移動のための Read アクセス（以下，階層間 Read という）より高い優先度で処理される．従って，仮に同一ドライブへの競合する Read があった場合のレスポンスタイムは，上位側 Read は概ね通常通りであり，階層間 Read は通常より大きくなると考えられる．

図 9 で読み取ったレスポンスタイムはこれらが平均化された値であり，FAST VP がない場合に比べてより大きな値であると考えられ，シミュレーション結果は FAST VP 構成のペナルティが大きいという見方ができる．一方で，階層間 Read のためにキャッシュへ読み込まれたデータが，上位側 Read によりアクセスされるケースも考えられ，この場合は FAST VP に有利な結果をもたらすと言える．

図 10 で読み取った IOPS では，FC と SATA が階層間 Read の影響を受けている [*7]．今回使用したワークロードとは異なる，授業のない閑散期に測定した FAST VP による階層間 Read の値によると，FC から SATA への移動は約 547IOPS，SATA から FC への移動は約 391IOPS であった．これらは今回使用したピーク時ワークロードの 15,675IOPS に対して 6%程度の値である．影響がないとは言えないが，前述の通り，階層間 Read によって上位側へのレスポンスにいい影響を与えていることも考えられる．

階層間 Read はレスポンスタイム及び IOPS の両方で好影響と悪影響を与えており，全体に対する影響の程度は比較的軽微であることから，本論文ではこれらの影響は相殺するものと考えることとした．厳密に判断するには，FAST VP によるデータ移動を停止した上で，今回と全く同じワークロードで測定し直して比較するしかなく，実運用データを用いた検証は非常に困難である．

第 2 は，全体の平均レスポンスタイムを求めるに当たって，VMAX エンジンのキャッシュによるレスポンスタイムが不明なため，近似値として SSD 階層における平均レスポンスタイムを用いた点である．この値は実際のキャッシュに対する応答とは関係しない値であり，キャッシュの応答はこれよりも高速であると考えられる．今回のシミュレーションでは，レスポンスタイムの合計時間における ReadMiss の割合が増えるほど FAST VP に有利な結果となるため，この近似のために FAST VP に有利な結論とはなっていない．

6. 関連研究

Shikida ら [10] は，24TB の SSD と多数のニアライン SAS ディスクを用いて 3PB 以上を実装した，Dell 社製 EqualLogic による自動階層化ストレージの設計と運用について報告している．ドライブの階層が 2 層であること，ECCS2012 におけるホームサービスに相当する機能のみを提供していることなどが異なるが，アクセスのないデータを自動的に低速な階層へ移動させるという機能は FAST VP と同じである．性能については SPECsfs2008_nfs.v3 によるベンチマーク結果を示しているが，実際のワークロードに基づく分析は提示されていない．Shikida らは自動階層化のポリシーとして，どの程度アクセスがなければ低速ドライブへ移動させるかという期間を設定している．ECCS2012 では，ブロックのアクセス頻度を学習する期間や，データ移動を積極的に行うかどうかの指標を調整している．いずれの設定方法についても，結果として生じる各階層の割り当て容量やドライブの負荷，レスポンスタイム

[*7] SSD から下位の階層へ移動する傾向はほとんど見られず，十分無視できる値である．

表 3 ワークロード全体に対するレスポンスタイム
Table 3 Response time of the whole workload.

種別	IOPS	FAST VP		構成 1（FC 15,000rpm）		構成 2（FC 10,000rpm）	
		レスポンス	合計時間	レスポンス	合計時間	レスポンス	合計時間
ReadHit+SeqRead	4,311	2	8,622	2	8,622	2	8,622
ReadMiss	3,307	7.3	24,141	12	39,684	24.3	80,360
Write	8,057	2	16,114	2	16,114	2	16,114
Total	15,675	—	48,877	—	64,420	—	105,096
平均レスポンスタイム			3.12		4.11		6.70
比			1		1.32		2.15

の変化を追跡して検証することが必要であり，チューニングには長期間のデータ収集と分析が重要である．

近年，データの自動階層化は多くの製品に搭載されている．Symmetrix VMAX は VMAX エンジンでブロック単位による学習を行っているが，F5 社製 ARX のように NAS ゲートウェイ [*8] として機能し，ポリシーを設定することでファイル単位で再配置の制御を行うものもある．いずれが適しているかはストレージの用途により異なるが，VMAX のようにストレージ自体が自動階層化の機能を持っている構成の方が管理対象の機器の点数が少なくなり，管理コストを抑制できると言える．

ECCS2012 ではドライブの空き容量をホームサービスとメールサービスで共用するため，単一のストレージプールとして集約する機能を重視した．必要に応じてファイルシステムを伸長するという観点では，EMC 社製 Isilon といった NAS 製品も市場に登場している．ECCS2012 の場合，利用者の増加を予測しにくいメールサービスに適した機能と言えるが，比較的少数のサーバから高負荷がかかるという利用形態で性能を発揮するかは，検証が必要である．

7. まとめ

本論文では，以下 2 つの機能に着目したストレージの設計及び構成について述べた：(1) SSD/FC/SATA のドライブを混在させ，アクセス頻度に応じてデータを自動再配置する機能，(2) これらドライブを単一のストレージプールに統合して，複数のサービスで空き容量を共有する機能．また，(1) の機能を用いることで，FC ディスクのみを用いた一般的な構成に比べて，レスポンスタイムが約 0.76 倍短縮されることをシミュレーションにより示した．

データ収集対象とした期間より前には，想定よりも FC の階層が使われず，割り当て・負荷とも SATA に偏った状況が見られたため，FAST VP の機能を含むストレージのファームウェア更新やパラメータの変更など，試行錯誤を重ねてきた．VMAX を用いることで詳細なレスポンスタイム等の指標を得られるようになったため，データ収集を継続し，次期システムの設計に活かすことが必要である．

謝辞 本システムの構築及び運用にご協力いただいている日本電気株式会社の方々に感謝します．

[*8] 複数の NAS のフロントエンドに設置し，NAS プロトコルのルーティングを行う装置．

参考文献

[1] 丸山一貴，関谷貴之，妹川竜雄，和田佳久：教育用計算機システムにおけるエージェント方式によるデュアルブート端末管理，インターネットと運用技術シンポジウム 2012(IOTS2012)，pp. 39–46 (2012).
[2] Maruyama, K. and Sekiya, T.: ECCS2012 makes PCs and printers in computer labs accessible from off-campus environment, *Proceedings of the 2013 ACM Annual Conference on Special Interest Group on University and College Computing Services*, pp. 105–108 (2013).
[3] 安部達巳，田中哲朗，関谷貴之，丸山一貴，前田光教，有賀　浩：教育用計算機ユーザ管理システムの改善と運用評価，大学 ICT 推進協議会 2012 年度年次大会論文集，pp. 277–281 (2012).
[4] 丸山一貴，関谷貴之：学外システム連携による教育用計算機システムプリントサービス，情報処理学会研究報告，Vol. 2012-IOT-16, No. 9 (2012).
[5] 丸山一貴，紙谷哲史，関谷貴之：学外システム連携による教育用計算機システムプリントサービスの利用動向，マルチメディア，分散，協調とモバイル (DICOMO2014) シンポジウム，pp. 1271–1277 (2014).
[6] Corporation, E.: *IMPLEMENTING FAST VP AND STORAGE TIERING FOR ORACLE DATABASE 11g AND EMC SYMMETRIX VMAX* (2011).
[7] Benenati-Romano, K., Otte, R. and Fried-Tanzer, D.: *Mainframe EMC Symmetrix Remote Data Facility (SRDF) Four-Site Migration* (2010).
[8] EMC Education Services[著](株) クイープ [訳]：IT 技術者なら知っておきたいストレージの原則と技術，インプレスジャパン (2013).
[9] EMC ジャパン株式会社：EMC SYMMETRIX VMAX ストレージシステムスペックシート（オンライン），入手先 〈http://japan.emc.com/collateral/hardware/specification-sheet/h6176-symmetrix-vmax-storage-system.pdf〉（参照 2014/09/05）.
[10] Shikida, M., Nakano, H., Kozaka, S., Mato, M. and Uda, S.: A Centralized Storage System with Automated Data Tiering for Private Cloud Environment, *Proceedings of the 2013 ACM Annual Conference on Special Interest Group on University and College Computing Services*, pp. 1–5 (2013).